NLSY97 GEOCODE CODEBOOK SUPPLEMENT

ROUND 3

Introduction to the NLSY97 Geocode Data

The NLSY97 geocode file contains detailed information on the geographic residence of each NLSY97 respondent. These data permit researchers to analyze the influences that a respondent's area of residence or type of environment may have on outcomes such as education and employment. In addition to the geographic data, the geocode CD-ROM includes all the variables on the main and event history NLSY97 CD-ROMs.

In addition to variables identifying the location of the respondent's residence, this supplemental data file contains selected variables drawn from other data sources and matched to the respondents' residences or schools. The majority of these variables are taken from the *County and City Data Books*; other outside sources include the BLS publication *Employment and Earnings* and the Integrated Postsecondary Education Data System (IPEDS). This introduction provides a discussion of the creation of the variables available on the NLSY97 geocode data file.

Geocode Data File Creation Procedures

The software package Maptitude (V4.2) is used in the creation of the NLSY97 geocode data file. This program links respondent address data to standard geographic information such as the FIPS (Federal Information Processing Standards) codes for state and county. Three graduated matching methods were applied, depending on the quality of the address data available.

1. An automated match was done between the respondent's locating address data and the Maptitude database. Address records with matching street segments were assigned the latitude and longitude of the location. In some cases, addresses had to be cleaned before they could be matched by the Maptitude program. Cleaning involves steps such as standardizing the address format, correcting obvious misspellings, identifying apartment numbers and locating them in the correct field, etc. It does not include any changes that might result in a change in the actual address location.
2. For some addresses, the procedure outlined in Step #1 failed to produce a match between the respondent's address data and the Maptitude database. In these cases, geocode staff used the Maptitude program to locate the correct street. If the street number could be located along this street, the latitude and longitude were assigned. However, some streets in the Maptitude database do not include information about street numbers. If this is the case, the address is manually located in the center of the street. The street is then classified as either a short street or a long street. Long streets cross Census tract or block group boundaries while short streets do not. As a result, the level of certainty about geographical information is much higher for short streets than for long streets.
3. Addresses unmatched by either of the first two procedures were assigned latitude and longitude coordinates according to a 5-digit zip centroid. A centroid is essentially the midpoint of a zip code area. The geographic information is less certain for respondents located using the zip centroid method.

Researchers can identify the method used to locate the respondent's address by using the variable GEO06, which provides information about the quality of the geographic match. This variable differentiates between addresses located based on the actual address, in the center of a short street, in the center of a long street, or using the zip centroid method. This variable can be used to determine the level of certainty for the respondent's geographic data.

Key variables describing the location of each respondent's residence are created by the Maptitude program using the longitude and latitude information. These two variables, question names GEO01 and GEO02, provide the FIPS code for the respondent's county and state of residence. FIPS codes are listed in Attachment 100 in this document.

The respondent's residence is further described by Metropolitan Statistical Area (MSA), Consolidated Metropolitan Statistical Area (CMSA), or New England Consolidated Metropolitan Area (NECMA). The MSA, CMSA, and NECMA definitions in the 1994 *County and City Data Book (CCDB)* are used to create the MSA variable. Survey staff compare the respondent's county and state of residence to the MSA definitions to determine if the respondent resides in an MSA. If so, GEO03 lists the respondent's MSA, CMSA, or NECMA of residence; GEO04 lists the Primary Metropolitan Statistical Area (PMSA) code for respondents residing in a CMSA (see Attachment 101 in this publication for further explanation). To assist researchers in easy classification of the MSA variable, GEO5 reports whether the code in GEO3 is an MSA, CMSA, or NECMA. Users should note that there are some slight differences between the 1994 codes and the Census Bureau's FIPS standards, as described in Attachment 101 in this document.

User Notes: Researchers should be aware that the geocoding procedure has been changed since earlier rounds of the NLSY97. All geocode variables from all rounds have been recreated using the new, more accurate method. Regardless of the round of interest, researchers should use data from the round 3 geocode CD.

Supplementary Created Variables

At this point, survey staff first created three geographic variables that are released with the main file NLSY97 data and also included on the geocode CD. First, CV_CENSUS_REGION provides researchers with information about whether the respondent lives in the Northeast, North Central, South, or West region of the country as defined by the Census Bureau. The states comprising each region are listed in the codebook for the region variable as well as section 4.6.3 of the *NLSY97 User's Guide*.

The second variable, CV_MSA, identifies whether the respondent lives in the central city of an MSA, in an MSA but not in the central city, or outside of an MSA. The central city boundaries are defined by the 1992 Census TIGER/Line files and are included in the Maptitude program; the MSA definitions used in this variable are the standard Census Bureau definitions rather than those drawn from the 1994 *CCDB*. This means that a few respondents may be listed as residing in an MSA in the status variable but do not have an MSA code in GEO03. Attachment 101 lists the differences between the 1998 Census and 1994 *CCDB* metropolitan statistical areas.

Finally, the main file variable CV_URBAN-RURAL indicates whether the respondent lives in an urban or rural place. Places are identified as urban or rural by the Census Bureau. Urban places are those in "urbanized areas" or "places" with a population of at least 2,500; all other areas are rural. According to the Census Bureau, about 25 percent of the U.S. population lives in rural areas.

If the respondent's residence was located using a street name match (method 2 above) or a zip centroid match (method 3), the MSA status and urban/rural variables are further evaluated. For the MSA variable, if the street or zip code falls completely inside or outside the boundaries of the central city, then the respondent is assigned to the appropriate status. If the street or zip code crosses the boundaries of the central city, then the respondent is coded as living in an MSA, with central city status unknown.

Similarly, respondents are only assigned to an urban or rural status if their entire street or zip code lies within an urban or rural area. If the street or zip code crosses an urban/rural boundary, the respondent is assigned to an unknown status.

A number of additional created variables are provided in the NLSY97 geocode data file for the convenience of researchers. In rounds 1–3, the geographic information was presented at the county level. Unless otherwise noted, these variables are based on the 1994 *County and City Data Book (CCDB)*, prepared by the U.S. Bureau of the Census. The *CCDB* data file includes information from the 1990 Census of the Population and Housing, the Current Population Surveys, and other supplemental data derived from a variety of federal government and private agencies. Table 1 lists the *CCDB* variables included on the geocode CD.

TABLE 1. County-level geocode variables

GEO7	1990 land area in square miles	GEO41	1988 active nonfederal physicians per 100,000 population [copyright]
GEO8	1992 population	GEO42	1991 community hospital beds per 100,000 population [copyright]
GEO9	1992 population, 1990 square miles	GEO43	1991 serious crimes per 100,000 population
GEO10	1990 population by race, White	GEO44	1990 persons 25 years and over, % high school graduate or higher
GEO11	1990 population by race, Black	GEO45	1990 persons 25 years and over, % with bachelor's degree or higher
GEO12	1990 population by race, American Indian, Eskimo, or Aleut	GEO46	1989 median family money income
GEO13 GEO14	1990 population by race, Asian or Pacific Islander	GEO47	1989 per capita money income
GEO15	1990 Hispanic origin population (of any race)	GEO48	1989 % of families with income below poverty level
GEO16	1990 Hispanic origin population, % of total population	GEO49	1990 total families—base for GEO48
GEO17	1990 population by age, % under 5 years	GEO50	1989 % of families with female householder (no spouse present) below poverty level
GEO18	1990 population by age, % 5 to 17 years	GEO51	1990 female householders (no spouse present), family households—base for GEO50
GEO19	1990 population by age, % 18 to 20 years	GEO52	1989 % of persons with income below poverty level
GEO20	1990 population by age, % 21 to 24 years	GEO53	1989 % of related children under 18 years below poverty level
GEO21	1990 population by age, % 25 to 34 years	GEO54	1990 workers 16 years and over, % working outside county of residence
GEO22	1990 population by age, % 35 to 44 years	GEO55	1991 civilian labor force
GEO23	1990 population by age, % 45 to 54 years	GEO56	1991 civilian labor force—number unemployed
GEO24	1990 population by age, % 55 to 64 years	GEO57	1991 civilian labor force—unemployment rate
GEO25	1990 population by age, % 65 to 74 years	GEO58	1990 civilian labor force
GEO26	1990 population by age, % 75 years and older	GEO59	1990 civilian labor force—% female
GEO27	1990 population—base for GEO15 to GEO25	GEO60	1990 civilian labor force—% unemployed
GEO28	1990 male population	GEO61	1990 civilian labor force—number employed
GEO29	1990 % of persons 5 years and older living in different house in 1985	GEO62	1990 civilian labor force, % employed in agriculture, forestry, and fisheries
GEO30	1990 % of persons 5 years and older living in different house, same state in 1985	GEO63	1990 civilian labor force, % employed in manufacturing
GEO31	1990 % of persons 5 years and older living in different house, different state in 1985	GEO64	1990 civilian labor force, % employed in wholesale and retail trade
GEO32	1990 family households, percent with own children under 18 years	GEO65	1990 civilian labor force, % employed in finance, insurance, and real estate
GEO33	1990 female householders (no spouse present), family households	GEO66	1990 civilian labor force, % employed in health services
GEO34	1990 female householders (no spouse present), family households, % with own child	GEO67	1990 civilian labor force, % employed in public administration
GEO35	1988 number of births	GEO68	1990 per capita personal income
GEO36	1988 births, % to mothers under 20 years		
GEO37	1988 births per 1,000 population		
GEO38	1988 population—base for GEO 36 and GEO39		
GEO39	1988 number of deaths		
GEO40	1988 deaths per 1,000 population		
	1988 deaths of infants under 1 year per 1,000 live births		

Integrated Postsecondary Education Data System (IPEDS) Codes. During the interview, respondents report the name and location of each college they have attended. Information from the Integrated Postsecondary Education Data System (IPEDS) is then merged with the name and address of the youth's college to provide users with the code identifying the school (GEO69) and its location (GEO70). More information about college codes is provided in Attachment 102.

Unemployment Rate Variable Creation. The next supplemental created variable, GEO71, is the unemployment rate for the respondent's area of residence. Unlike the unemployment variables listed in

Table 1, which are reported for the respondent's county, this variable provides the unemployment rate for the respondent's metropolitan area, if applicable, or the balance of the respondent's state. The round 1 NLSY97 unemployment rate variable was constructed using state and metropolitan area labor force data from the May 1998 publication of *Employment and Earnings* for the month of March 1998. The round 2 data were taken from the May 1999 publication for the month of March 1999, and the round 3 variable was based on the June 2000 publication for the month of March 2000. *Employment and Earnings*, published by the U.S. Department of Labor, Bureau of Labor Statistics, lists the civilian labor force and number of unemployed persons for every state and metropolitan area.

The respondent's metropolitan statistical area is assigned based on the county, state, and zip code of current residence using the process described earlier in this introduction. If a respondent resides in a metropolitan area that is listed in *Employment and Earnings*, then the unemployment rate in the NLSY97 variable is the unemployment rate for that metropolitan area. This rate is calculated by dividing the number of unemployed persons by the number of people in the civilian labor force as reported by BLS.

If the respondent does not reside in a metropolitan area, he or she is assigned a "balance of state" unemployment rate. For these cases, the figures provided for the state and its metropolitan areas are used to compute the unemployment rate for the portion of the state that is not represented in any metropolitan statistical area. Because the *Employment and Earnings* numbers are based on an older set of MSA codes than the NLSY97 data, there are also some cases in which NLSY97 metropolitan areas do not match those used in *Employment and Earnings*. In these cases, respondents are assigned the balance of state unemployment rate even though they do reside in a metropolitan area. (Interested users can examine the *Employment and Earnings* MSA definitions in each year's May edition; NLSY97 codes are provided in Attachment 101 in this document.)

Randomized Code for Primary Sampling Unit (PSU). Finally, the geocode CD includes a variable identifying respondents who lived in the same sampling area in the initial survey year. Defined by NORC, primary sampling units are the areas used to draw the sample for the NLSY97 (see chapter 2 of the *NLSY97 User's Guide* for details on sampling procedures). This variable, GEO72, presents a scrambled version of the PSU code so that researchers can identify respondents from the same area but cannot determine the exact PSU from which they were drawn. The randomized PSU data can be used in the estimation of design effects for the NLSY97 sample.

Migration History Variables

In the household information (YHHI) section, rounds 2 and 3 of the NLSY97 survey collected information about each residence of the respondent since the previous interview date. Respondents who had moved to a different city, county, or state were asked to report the date of the move and the new city, county, and state of residence. These data were recorded for each move. In round 3, these data were geocoded using the standard state and county FIPS codes. The codes are included on the geocode CD so that researchers can track respondents' migration patterns. Data collected in the round 2 interview are planned for inclusion on the geocode file but may not be available until a future data release; researchers should contact NLS User Services for more information.

Geocoded migration variables can be located in the data set by searching for question names that start with "GEO_M." For each move, the respondent will have a state variable (e.g., GEO_M_ST.01) and a county variable (e.g., GEO_M_CO.01). The number at the end of the question name indicates which move the data apply to. For example, variables ending in ".01" refer to the first move after the last interview, those ending in ".02" refer to the second move, and so on.

Birth Date Variables

Data indicating the day of birth of the respondent, his or her parents, children, and other household members are included on the geocode CD-ROM. Month and year of birth variables appear in the main public use data set. The reference numbers and question names for the day of birth variables correspond to those used in the main data set for month and year of birth. For example, variables KEY!BDATE_M and KEY!BDATE_Y (R05364.01 and R05364.02) in the main data set contain information about the respondent's birth month and year. The corresponding variable in the geocode data is KEY!BDATE_D (R05364.00) and provides information about the respondent's day of birth.

User Notes: Due to technical considerations, the day of birth variables in the geocode data do **not** have the word "day" in the title, nor do they show information about day of birth in the codebook. Instead, these variables have "month/year" in the title and list month and year information on the codebook page. However, the actual data do reflect day of birth information. Researchers can verify that they are using day of birth data by examining the question name, which contains a "D" for day variables.

Missing Data

Missing values for geocode variables are assigned using the same coding system as the main NLSY97 data file; see the *NLSY97 User's Guide* for more information. In general, the following codes have been used:

–3 = Invalid skip. Address data cannot be geocoded. In these cases, the respondent lives in the United States but has provided incomplete or conflicting address information. The data file contains as much information as possible for these respondents; for example, if survey staff are confident that the state is correct but cannot identify the county, the state variable will have a valid code and the county variable will have a value of –3.

–4 = Valid skip. The respondent has no information for this variable because it does not apply. For example, respondents living outside the United States have a –4 for all residence variables; respondents not attending college have a –4 for the college variables.

–5 = Noninterview. In round 2 and beyond, respondents who did not participate in the survey are assigned a value of –5 for all data in that round.

Use of the File

Finally, we have a few suggestions concerning the use of these NLSY97 geographic data files. First, the data file and the accompanying documentation should be used in conjunction with the printed version of the 1994 *County and City Data Book* and the IPEDS codes so that researchers have complete information regarding variable descriptions and coding idiosyncrasies. Second, the data should not be used in any fashion that would endanger the confidentiality of any sample member. To use these data, users must sign a written licensing agreement consenting to protect respondent confidentiality and to other conditions; agree not to make, or allow to be made, unauthorized copies of the geocode file; and further agree to indemnify the Center for Human Resource Research for all claims arising from misuse of the file.